计算机科学
COMPUTER SCIENCE

基于RoBERTa和加权图卷积网络的中文地质实体关系抽取

张鲁，段友祥，刘娟，陆誉翕

引用本文

张鲁，段友祥，刘娟，陆誉翕．基于RoBERTa和加权图卷积网络的中文地质实体关系抽取[J]．计算机科学，2024，51(8)：297-303.

ZHANG Lu, DUAN Youxiang, LIU Juan, LU Yuxi. Chinese Geological Entity Relation Extraction Based on RoBERTa and Weighted Graph Convolutional Networks [J]. Computer Science, 2024, 51(8): 297-303.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230600231

基于 RoBERTa 和加权图卷积网络的中文地质实体关系抽取

张 鲁 段友祥 刘 娟 陆誉翕
中国石油大学(华东)计算机科学与技术学院 山东 青岛 266580
(s21070043@s.upc.edu.cn)

摘 要 知识是大数据和人工智能的基石,知识图谱的可解释性和可扩展性等优势使其成为智能系统的重要技术。智能决策在各个领域都有迫切的应用需求,为知识图谱提供基于数据分析和推理的决策支持和应用场景,但领域场景复杂、数据多源、知识维度广,因此知识图谱的构建和应用都面临着很多挑战。针对地质领域知识图谱构建过程中领域知识模式完备性差的问题,以及现有实体关系抽取方法在处理非欧氏数据时存在的不足,提出了一种基于图结构的实体关系抽取模型 RoGCN-ATT。该模型使用 RoBERTa-wwm-ext-large 中文预训练模型作为序列编码器,结合 BiLSTM 获取更丰富的语义信息,使用加权图卷积网络结合注意力机制获取结构依赖信息,以增强模型对关系三元组的抽取性能。在地质数据集上 F1 值达 78.56%,与其他模型的对比实验表明,RoGCN-ATT 有效提升了实体关系抽取性能,为地质知识图谱的构建和应用提供了有力的支持。

关键词:实体关系抽取;图卷积网络;依存句法分析;注意力机制;地质领域
中图分类号 TP391

Chinese Geological Entity Relation Extraction Based on RoBERTa and Weighted Graph Convolutional Networks

ZHANG Lu,DUAN Youxiang,LIU Juan and LU Yuxi
College of Computer Science and Technology,China University of Petroleum(East China),Qingdao,Shandong 266580,China

Abstract Knowledge is the cornerstone of big data and artificial intelligence. Knowledge graphs offer interpretability and scalability advantages, making them crucial in intelligent systems. Intelligent decision has urgent application demand in various fields, providing decision support and application scenarios for knowledge graphs based on data analysis and reasoning. However, constructing and applying knowledge graphs face challenges due to complex domain scenarios, multi-source data, and extensive knowledge dimensions. To address the problem of incomplete domain knowledge patterns during geological domain knowledge graph construction and the limitations of existing entity relationship extraction methods in dealing with non-Euclidean data, a graph structure-based entity relationship extraction model RoGCN-ATT is proposed. This model utilizes RoBERTa-wwm-ext-large, a Chinese pre-trained model, as the sequence encoder combined with BiLSTM to capture richer semantic information. It also employs weighted graph convolutional networks along with attention mechanisms to capture structural dependency information and enhance the extraction performance of relation triplets. Experimental results show that the F1 value reaches 78.56% on the geological dataset. Compared with other models, RoGCN-ATT effectively improves the entity-relationship extraction performance and provides strong support for the construction and application of geological knowledge maps.

Keywords Entity relation extraction, Graph convolutional networks, Dependency parsing, Attention mechanism, Geology domain

1 引言

以深度学习为代表的人工智能新技术已经成为地质领域研究和应用的重要技术支撑。随着新技术、新方法和新工艺的引入,一方面地质认知的数据呈爆炸式增长,数据形式日益多样,另一方面对地质大数据的分析、解释,特别是应用数据进行地质生产决策支持也提出了巨大挑战。因此,如何从地方海量、繁杂的数据中获取更多的决策支持需要的知识,并对知识进行细粒度描述是一个亟待解决的问题。

知识图谱作为一种结构化的知识表示方式,为人工智能系统提供了丰富的语义关联和知识资源。又因其能够赋予智能体分析、推理和理解等能力,被广泛应用于智能搜索、问答系统等任务,为领域决策支持提供了新的技术途径。知识图谱的质量直接影响着其下游应用,其核心技术包括知识图谱

到稿日期:2023-06-29 返修日期:2023-11-12
基金项目:中央高校基本科研业务费专项资金(20CX05017A);中石油重大科技项目(ZD2019-183-006)
This work was supported by the Fundamental Research Funds for the Central Universities of Ministry of Education of China(20CX05017A) and Major Scientific and Technological Projects of CNPC(ZD2019-183-006).
通信作者:段友祥(yxduan@upc.edu.cn)

构建、知识图谱存储和知识推理与计算等，其中知识图谱构建是基础，也是知识图谱质量的基本保障。而在构建的过程中，关系实体抽取是关键，目的是通过对实体和关系进行深度挖掘和分析，将现实世界的实体和实体之间的关系用结构化的形式表现出来，以帮助地质学家和研究人员更好地理解地质领域知识。此外，领域知识由于其专业性的特点，知识的表示和获取更难，因此如何提高领域知识的完备性是一个具有挑战性的难题。

知识图谱构建的关键是知识抽取，这也是知识图谱研究的难点。实体关系抽取技术是知识抽取任务的重要研究方向之一，从非结构化文本中挖掘实体间的语义关系构建三元组，为知识图谱的建立奠定了基础[2]。传统的实体关系抽取方法依赖于人工标注，且专注于特定领域。近年来，随着深度学习和自然语言处理技术的发展，基于神经网络的实体关系抽取成为主流。目前，实体关系抽取任务主流方法主要分为两类：1)流水线关系抽取方法[3-5]，即先进行实体识别任务再进行关系抽取任务；2)实体关系联合抽取方法[6-9]，即两个任务同时进行。流水线关系抽取方法相比传统的机器学习方法，其优势在于灵活性更强、可解释性强，但容易产生误差。而实体关系联合抽取方法被认为具有更好的性能和潜力，解决了流水线方法中的错误传播问题，但是需要更复杂的结构编码来获取更丰富的信息[10]。尽管实体关系抽取任务的两类主流方法的表现都比较出色，但对于非欧氏数据的处理仍存在问题和挑战，而这一问题在地质等领域知识工程中更为突出。目前，地质领域内中文文本实体关系抽取研究还处于起步阶段，相关数据集和模型也相对较少，因此如何有效地将深度学习和知识图谱等技术应用到地质学实体关系抽取中，仍然需要进一步的研究和探索。

本文采用 RoBERTa-wwm-ext-large 中文预训练模型[11]（以下简称 RoBERTa 中文预训练模型）进行嵌入表示，并结合双向长短期记忆网络（Bidirectional Long Short-Term Memory，BiLSTM）[12]进行序列编码，使用图卷积神经网络和注意力机制分析句子中各个成分的依赖关系，提出了结合注意力机制的图卷积网络模型 RoGCN-ATT（RoBERTa Graph Convolutional Network Attention），从而为实体关系抽取提供有效的结构化信息，帮助模型整理句子结构，提升实体关系抽取性能。

2 相关工作

传统的实体关系抽取通常需要人力参与进行特征提取，为减少人力资源，基于神经网络的实体关系抽取方法获得了研究者的关注。Kumar[13]在对用于关系抽取的深度学习模型的回顾中指出，基于序列特征的模型如卷积神经网络（Convolutional Neural Network，CNN）[14]和循环神经网络（Recurrent Neural Network，RNN）[15]都能为文本分类任务提供互补信息。Nasar 等[16]指出目前主流的命名实体识别（Named Entity Recognition，NER）方法为采用结合条件随机场（Conditional Random Field，CRF）的 CNN 和 RNN 的混合模型，并在大多数基准数据集上都取得了最先进的结果。在地质领域，针对数据中的噪声和数据稀疏性问题，Lei 等[17]基于具有层次结构的领域词汇表，构建了一个深度置信网络，并结合注意力机制来提高了地质灾害实体识别的准确率。Fan 等[18]结合了一个多分支双向门控递归单元（Bidirectional Gated Recurrent Unit，BiGRU）层和一个 CRF 模型用于 NER，并基于该模型构建了一个大规模的地质灾害知识库。

对于关系抽取问题，目前广泛使用远距离监督方法和基于深度学习的联合模型来解决。Luo 等[19]引入高速网结合 BGRU 网络提出了 Att-BGRU-HN 模型，并将其应用于地质数据分析，提升了关系抽取的性能。Huang 等[20]基于预训练模型 BERT 提出了一个远程监督关系抽取模型，用于对金矿地质特征进行提取。Chen 等[21]在 BERT 的基础上结合 BiLSTM 提取上下文信息，提出了 BERT-BiLSTM-CRF 模型用于提取岩石三元组，F1 值达 91.75%。Wang 等[22]基于句法结构建立了开放式实体关系联合抽取模型 CSSEM，在中文地质数据集和通用数据集上证明了该模型的有效性。针对地质数据集中的重叠关系，Wu 等[23]结合 BERT-wwm 模型，提出了层级标注模型 HtERT，用于处理重叠关系三元组，更精确地表示了地质样本信息。

为了更好地捕捉实体之间的关系，基于图结构的方法逐渐被广泛应用于实体关系抽取任务中。

基于图结构的实体关系抽取方法是将实体看作节点，将实体之间的依存关系看作边，从而构成图结构。Bunescu 等[24]首次观察到，在图结构中，与实体关系抽取相关的信息几乎完全集中在两个实体之间的最短依存路径（Shortest Dependency Path，SDP）中，许多研究者基于此开始进行了细化和改进。Cai 等[25]利用 SDP 的依赖信息，结合 CNN 和 BiLSTM 提出了 RCNN 模型，在公共数据集 SemEval-2010 Task 8[26]（以下简称 SemEval2010）上 F1 值达到了 82.4%，大大提升了关系抽取的性能。为了降低冗余信息的影响，众多研究者开始采用图卷积网络（Graph Convolutional Network，GCN）[27]方法。Yu 等[28]使用软剪枝策略，引入反思机制结合 GCN，不依赖预定义的规则实现了剪枝依赖树并将其用于关系抽取，在 SemEval2010 数据集上相比 RCNN 模型 F1 值提高了 4%。

在对图结构信息进行编码时，为了有效表达节点信息，Hong 等[29]结合注意力网络构建完全图，采用实体跨度表示节点，边为关系表示，充分利用相邻节点特征信息。除了在单个句子中的应用，基于图结构的实体关系抽取也被广泛应用于上下文关系抽取。Tian 等[30]利用注意力机制结合 GCN，提出了一种依赖驱动的模型 A-GCN+BERT，在 SemEval2010 数据集上 F1 值达 89.85%。Zhou 等[31]将 GCN 用于文档级关系抽取，以捕捉实体的全局上下文信息。Duan 等[32]为增强句间信息的推理能力，提出了一种图注意力卷积网络模型，用于文档级的关系抽取。

目前，RIFRE[33]作为基于图结构的实体关系抽取 SOTA 模型，基于异构图神经网络将关系和词建模为图上的节点，并使用消息传递机制反复融合这两类语义节点，在节点更新后进行关系抽取，在 SemEval2010 数据集上相比 A-GCN+BERT 模型 F1 值提高了 1.45%，相比 RCNN 模型 F1 值提高了 8.9%。

3　本文方法

由于地质数据的非结构化及语义的多样性，在进行地质领域知识图谱构建时，同一个实体可能以不同的方式描述，增加了实体关系抽取的难度。而 GCN 作为一种基于图结构的深度学习方法，可以很好地用于这种复杂结构的处理。受 Zhou 等[34]逻辑邻接矩阵（Logical Adjacency Matrix，LAM）的启发，利用 RoBERTa 中文预训练模型对中文地质文本进行嵌入表示，并结合 GCN 引入注意力机制构建逻辑邻接矩阵，通过分类网络得到文本的标签预测结果，使模型更加适用于地质领域的实体关系抽取任务。RoGCN-ATT 的主要结构如图 1 所示，其中包括 3 个部分，即序列编码模块、WGCN 编码模块、表示融合和关系分类模块。

该模型解决的主要问题及创新性体现在：1）利用基于图结构的方法解决了对非欧氏数据的实体关系抽取；2）序列嵌入模块引入 RoBERTa 中文预训练模型对序列进行编码；3）WGCN 编码模块引入注意力机制表示权重构建逻辑邻接矩阵。

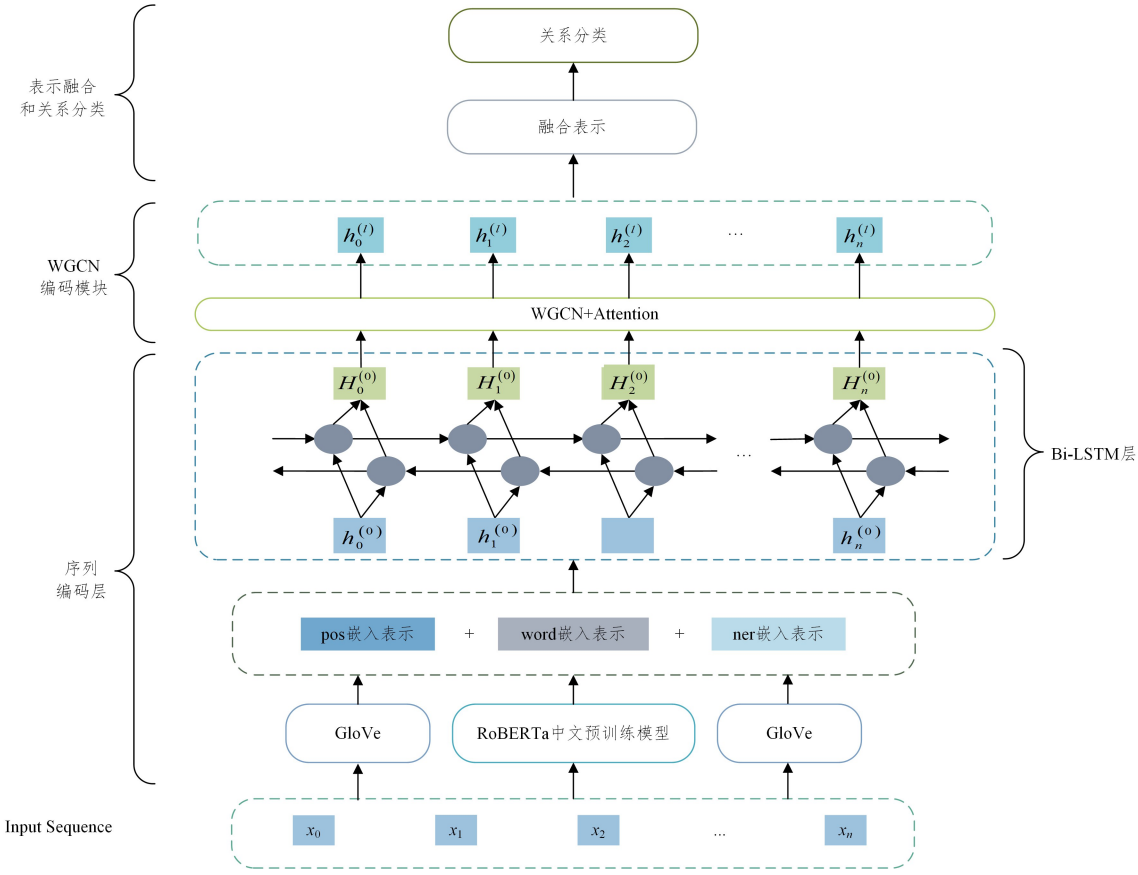

图 1　RoGCN-ATT 模型的框架

Fig. 1　Framework of RoGCN-ATT model

3.1　序列编码模块

在实体关系抽取任务中，序列编码模块作为常用模块，它主要用于将输入的文本序列转换为向量表示，以便于后续的处理和分析。RoBERTa 中文预训练模型相比 BERT 模型具有更强的语义理解和表征能力，因此本文中该模块主要使用 RoBERTa 中文预训练模型进行嵌入编码，BiLSTM 模型用于进一步编码。使用 RoBERTa 中文预训练模型进行词嵌入（word 嵌入），输入序列为 $x=[x_1,x_2,\cdots,x_n]$，其中 x_i 为序列的第 i 个词语。使用 RoBERTa 中文预训练模型将序列 x 转化为输入编码 $\boldsymbol{e}=[\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_n]$，其中 $\boldsymbol{e}_i$ 表示第 i 个词语的嵌入编码，通过 RoBERTa 中文预训练模型的前向传递计算得到序列表示，即 $\boldsymbol{h}=[\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_n]$，最终使用 LayerNorm 函数进行标准化处理，得到最终的 word 嵌入，如式(1)所示：

$$f(x)=\mathrm{LayerNorm}\left(\sum_{i=1}^{n}\alpha_i h_i\right) \tag{1}$$

其中，α_i 是一个标准化权重，计算式如式(2)所示：

$$\alpha_i=\frac{\exp(e_i^{\mathrm{T}}\boldsymbol{e}_c)}{\sum_{j=1}^{n}\exp(e_j^{\mathrm{T}}\boldsymbol{e}_c)} \tag{2}$$

其中，$\boldsymbol{e}_c$ 表示一个可学习的向量，作为一个查询向量来计算每个词语的注意力权重。使用 GloVe[35]对词性标签（pos）和实体识别标签（ner）进行嵌入表示，分别如式(3)和式(4)所示：

$$\boldsymbol{e}_{\mathrm{pos}}^{(0)}=GloVeEmb(x^{\mathrm{pos}})=[x_1^{\mathrm{pos}},x_2^{\mathrm{pos}},\cdots,x_n^{\mathrm{pos}}] \tag{3}$$

$$\boldsymbol{e}_{\mathrm{ner}}^{(0)}=GloVeEmb(x^{\mathrm{ner}})=[x_1^{\mathrm{ner}},x_2^{\mathrm{ner}},\cdots,x_n^{\mathrm{ner}}] \tag{4}$$

其中，x_i^{pos} 表示第 i 个单词的 pos 嵌入，x_i^{ner} 表示第 i 个单词的 ner 嵌入，将式(1)、式(3)、式(4)所示的向量拼接起来，得到最终的嵌入表示 $h^{(0)}$ 和维度 m，如式(5)和式(6)所示：

$$\boldsymbol{h}^{(0)}=[f(x);\boldsymbol{e}_{\mathrm{pos}}^{(0)};\boldsymbol{e}_{\mathrm{ner}}^{(0)}] \tag{5}$$

$$m=d_{\mathrm{word}}+d_{\mathrm{pos}}+d_{\mathrm{ner}} \tag{6}$$

其中，d_{word}，d_{pos}，d_{ner} 分别表示 word 嵌入、ner 嵌入和 pos 嵌入的维度，在拼接之后得到一个新的二维矩阵 $\boldsymbol{h}^{(0)}\in\mathbb{R}^{n\times m}$。由于该矩阵中包含的信息并不完整，因此结合 BiL-

STM 层从前向和后向两个方向上读取序列，利用状态向量捕捉上下文信息，得到整个序列的编码表示 $\boldsymbol{h}_i^{(0)}$ 并用作后续模型的输入。

3.2 WGCN 编码模块

该模块主要用于获得句子的依赖表示，提取更有效的特征表示，用于后续的实体关系分类。图卷积神经网络作为挖掘图数据结构特征的有力模型，每个节点的特征向量可以融合相邻节点的特征得到更丰富的语义信息，因此图卷积神经网络凭借其独特的图结构特性学习实体嵌入表示的优势被应用于实体关系抽取任务中。与图卷积神经网络在图像处理方面的不同是，对于图像，图卷积神经网络主要是利用像素点的矩阵信息，图结构相对简单；而在实体关系抽取任务中，图卷积神经网络主要是将输入数据中的实体和关系作为节点和边构成依存句法分析树，结构相对复杂。因此，该模块将依存句法分析和图卷积网络相结合，通过融合句子的句法结构和实体之间的关系信息来获得更全面的上下文表示。

3.2.1 依存句法分析

图卷积神经网络通过在图结构上进行信息传递和特征学习，能够更好地理解和分析实体之间的依存关系，实现更准确的依存句法分析。依存句法分析树展示的是文本分词之间的语法和语义关系。依存句法分析树的根节点通常是句子的核心单词，一般使用“root”表示，有且只有一个节点依赖于它。节点和节点之间的依存关系用一个有向弧表示，叫作依存弧，本文中依存弧的方向统一为由从属词指向支配词。在依存句法分析树中，依存关系主要分为两种：一种是修饰关系，表示一个单词修饰另一个单词；另一种是中心关系，表示一个单词是另一个单词的中心或核心。如图 2 所示，“煤窑沟是中华人民共和国新疆维吾尔自治区的一条河流，位于塔里木盆地”这句话中“是”即为依赖于根节点的节点；“中华人民共和国”和“新疆维吾尔自治区”有向弧表示的是后者修饰前者的修饰关系；“煤窑沟”和“是”之间是主语和谓语动词的依赖关系，表示的是中心关系。

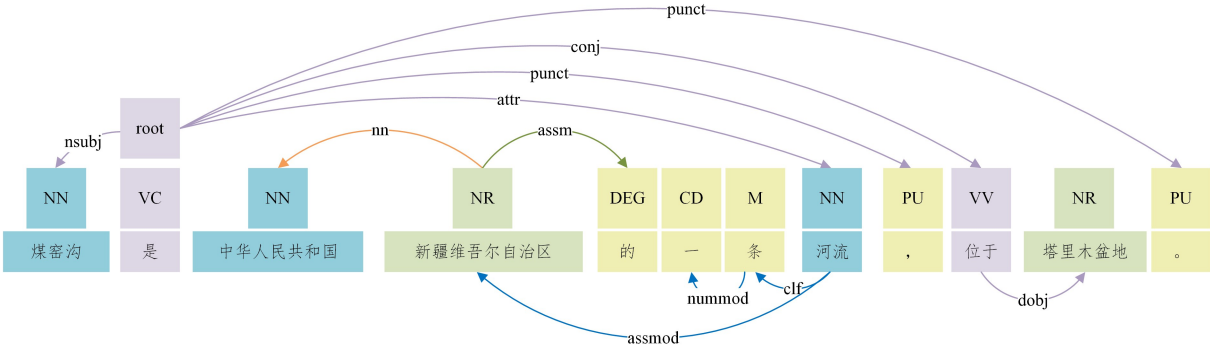

图 2 依存句法分析示例

Fig. 2 Example of dependency parsing

3.2.2 加权图卷积网络与注意力机制层

由于图卷积网络在经过多次卷积后会耗费大量时间并且导致较为严重的过平滑现象，因此引入加权图卷积网络(Weighted Graph Convolutional Network，WGCN)。加权图卷积网络是在图卷积网络的基础上进行的改进，对于输入的依存句法分析树，每个节点代表一个词语，边表示依存关系，通过添加虚拟边构造 LAM，用于获得 k 阶邻域的依赖关系，对于 LAM 中的权重值，采用节点间的相似度及距离表示，具体如式(7)、式(8)所示：

$$s_{ij}=\frac{\vec{x}_i \cdot \vec{x}_j}{\|\vec{x}_i\| \|\vec{x}_j\|} \tag{7}$$

$$d_{ij}=\frac{1}{e^{d-1}} \tag{8}$$

其中，s_{ij} 表示节点 i 和节点 j 的节点相似度，d_{ij} 表示两节点之间的距离的倒数，e 为欧拉数，使用 softmax 操作将两者结合进行归一化得到最终的注意力权重，具体如式(9)所示：

$$e_{ij}=\text{softmax}(s_{ij}, d_{ij}) \tag{9}$$

接着使用 e_{ij} 作为权重值来计算加权后的邻接矩阵，具体如式(10)所示：

$$\boldsymbol{A}_{ij}=\frac{\exp(e_{ij})}{\sum_{k\in N_i}\exp(e_{ik})} \tag{10}$$

其中，N_i 表示 i 节点的邻居节点集合，构建后的加权邻接矩阵如图 3 所示，其中初始化的邻接矩阵元素值均为 0，经过注意力机制权重计算后，绿色部分表示原序列中存在依赖关系的边的权重值，灰蓝色部分表示添加的虚拟边的权重值，最后将逻辑邻接矩阵与节点特征卷积得到最后的表征，如式(11)所示：

$$\boldsymbol{h}_i^{(l)}=\sigma\left(\frac{\sum_{j=1}^{n}\boldsymbol{A}_{ij}^{(l)}\boldsymbol{W}^{(l)}\boldsymbol{h}_j^{(l-1)}}{d_i}+\boldsymbol{b}^{(l)}\right) \tag{11}$$

其中，$\boldsymbol{h}_i^{(0)}$ 为序列编码模块使用 BiLSTM 层得到的最终向量表示 $\boldsymbol{h}_i^{(0)}$，σ 表示激活函数，d_i 为节点 i 在依存句法树中的度，$\boldsymbol{b}^{(l)}$ 为偏置向量。将相似度与距离结合作为权重值，并将其与逻辑邻接矩阵结合得到加权邻接矩阵，以确保节点之间的信息能够充分被捕捉，并保留不同节点之间的差异性。

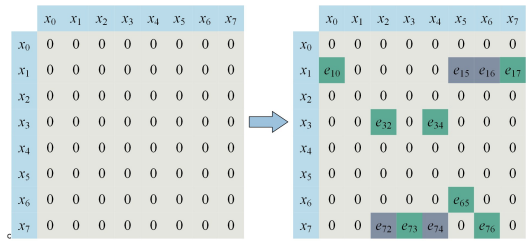

图 3 加权邻接矩阵图(电子版为彩图)

Fig. 3 Weighted adjacency matrix graph

3.3　融合表示和关系分类模块

该模块的目的是将 WGCN 编码模块得到的依存表示与实体表示进行有效的融合，以最终预测实体之间的关系类型。本文在该部分将句子和实体的所有表征融合到一起，然后根据输入的句子中提及的两个实体之间的关系进行分类，使用全连接层将包含序列信息和 WGCN 编码信息的最终融合表示 $\boldsymbol{h}_f$ 通过 softmax 操作传入一个前馈神经网络中，得到关系的概率分布，如式(12)－式(14)所示：

$$\boldsymbol{h}_f=\boldsymbol{h}_s;\boldsymbol{h}_i^{(l)};\boldsymbol{h}_o \tag{12}$$

$$\boldsymbol{p}=\mathrm{softmax}(\boldsymbol{h}_f)=\frac{\exp(\boldsymbol{h}_f)}{\sum_{j=1}^{k}\boldsymbol{h}_j} \tag{13}$$

$$\hat{y}=\underset{i\in k}{\mathrm{argmax}}(p_i) \tag{14}$$

其中，$\boldsymbol{h}_s$ 和 $\boldsymbol{h}_o$ 分别为主体客体的实体表示，$\boldsymbol{h}_i^{(l)}$ 为 WGCN 编码模块得到的依存信息，$\boldsymbol{p}\in\mathbb{R}^k$ 表示每个关系类别的概率，选择概率最大的关系作为两个实体之间的关系。使用交叉熵函数作为模型的损失函数，将模型输出的预测结果与实际关系类型进行比较，计算出模型预测结果与实际结果之间的误差，并将误差反向传播训练，以提高关系分类的准确性。

4　实验与分析

为了验证模型的有效性，本文与基于序列特征的模型、基于 BERT 的模型和基于图结构的模型进行了对比实验和消融实验。

4.1　数据集

为了更全面地评估本文模型在地质领域中的性能，使用的地质数据集是基于远程监督方法，根据中文地质词典中获取的地质实体在中文维基百科知识库中查找地质三元组，然后将三元组对齐中文维基百科语料库构建而来。在构建的数据集中，共标注了 9 种关系。为更好地体现实验的性能，我们从中选取了 7 种关系类型的数据，并从中选取了 30 238 条适用于本文模型的数据。其中随机抽取了数据中的 60%，20%，20%的数据分别作为训练集、验证集和测试集。该数据集中关系的类型及具体数量如表 1 所列。

表 1　地质数据集划分表示

Table 1　Partition of geological dataset representation

分类	训练集	测试集	验证集
子类	5410	2214	4417
实例	1993	512	596
位置	2893	2296	1591
属于	4174	1127	948
颜色	34	20	84
用途	483	18	11
other	4037	1648	1350

4.2　实验过程

本文实验服务器的相关配置为：GeForce RTX3060 显卡，显存为 12 GB；环境为 Python3.9，深度学习框架为 Pytorch1.11.0。

在输入文本嵌入层中，使用预训练的 RoBERTa 中文预训练模型对中文地质数据集进行句子序列嵌入，同时使用 GloVe 和 Shen 等[36]提供的维基百科词向量数据初始化句子中的 pos 嵌入和 ner 嵌入矩阵。此外，为了获得更多信息作为后续层的输入，使用 HanLP 工具[37]对句子进行处理，以获取分词结果、词性标注和依存关系等信息。

对于数据集，本文随机初始化 30d 的 ner 嵌入和 pos 嵌入。为了有效实现残差计算，将 BiLSTM 的隐藏大小设置为 100，将 WGCN 的隐藏大小设置为 200。通过调整参数，最终选择验证集中 F1 值最高的模型。与相关工作类似，本文使用实体关系抽取的精确率(Precision)、召回率(Recall)和 F1 值作为评估指标。具体参数设置如表 2 所列。

表 2　超参数设置

Table 2　Hyperparameter settings

参数	取值
RoBERTa embedding 维数	768
BiLSTM 编码维度	100
WGCN 层数	1
WGCN 隐藏层维度	200
Batch 大小	50
学习率	1.0
衰减率	0.9
epochs	150

4.3　结果与分析

4.3.1　对比实验结果分析

本文选取基于序列特征的经典模型 BiLSTM 与 PCNN＋ATT、基于 BERT 的模型 BERT-BiLSTM-CRF 与 HtERT 和基于图结构的经典模型 C-GCN 进行了对比实验。

1)BiLSTM[12]：双向长短时记忆网络挖掘句子级信息；

2)PCNN＋ATT[38]：引入注意力机制，结合 PCNN 缓解远程监督中的错误标签问题；

3)BERT-BiLSTM-CRF[21]：使用预训练模型 BERT 进行词嵌入，结合 BiLSTM 获取上下文信息，并使用 CRF 作为实体和关系之间转移规则的实体关系抽取模型；

4)HtERT[23]：基于 BERT-wwm 结合位置嵌入提取地质三元组；

5)C-GCN[39]：利用 SDP 并构造多个子图进行图卷积操作，以获得更多关系提取的有效信息。

以上 5 种模型都能对文本进行实体关系抽取，因此本文使用以上 5 种模型与本文模型 RoGCN-ATT 在 4.1 节中介绍的地质数据集上进行对比实验，具体结果如表 3 所列。

由于远程监督的特性，噪声数据会影响召回率并导致精确率偏低。从表中可以看出，基于序列标注的模型由于词与词之间的相关性效果较差，精确率相对较低。而基于图结构的模型使用图卷积网络可以更好地捕捉词与词之间的复杂关系，从而利用更多的特征信息，精确率相对较高。本文提出的模型 RoGCN-ATT 采用 RoBERTa 中文预训练模型结合 BiLSTM 模型提取语义信息，并使用注意力机制作为加权图卷积网络邻接矩阵的权重值获取结构特征，精确率和 F1 值都有明显提高，相比序列模型 HtERT，其精确率提高 2.65%，F1 值提高了 2.54%，相比图结构模型 C-GCN，其精确率提高了 15.16%，F1 值提高了 14.27%，这表明其在应用中可以有效地从复杂文本中提取实体关系信息。

表 3 对比实验结果

Table 3 Comparative experimental results

(%)

模型	*Precision*	*Recall*	*F*1
BiLSTM	55.35	57.43	56.37
PCNN+ATT	61.56	65.52	63.48
C-GCN	64.69	62.51	64.29
BERT-BiLSTM-CRF	76.55	73.88	75.65
HtERT	77.20	74.87	76.02
RoGCN-ATT	**79.85**	**76.05**	**78.56**

4.3.2 消融实验结果分析

针对本文模型，主要从以下 3 个方面进行了消融实验，以验证其有效性并探究不同因素对模型性能的影响。

1)依存句法分析工具的比较，实验名称设置为-Stanza。使用 Stanza[40]和 HanLP 实验效果对比，由于开发机制和语言等的影响，实验处理结果表明在处理中文数据方面 HanLP 比 Stanza 的精确率更高。

2)序列嵌入模型的比较，实验名称设置为-GloVe。使用 GloVe 和 RoBERTa 中文预训练模型的嵌入实验效果对比，GloVe 只是基于全局词向量进行嵌入，而 RoBERTa 中文预训练模型能够获得更丰富的语义信息，因此模型的 F1 值得到了有效提升。

3)邻接矩阵的设置，实验名称设置为-no-ATT。基线模型采用的是使用距离作为逻辑邻接矩阵中的权重值，本文模型采用注意力机制作为邻接矩阵中的权重值，增强了模型的表达能力，提高了模型的泛化能力。具体实验结果展示如表 4 所列，实验结果表明本文模型在中文地质数据集上表现出了良好的性能。

表 4 消融实验结果

Table 4 Results of ablation experiment

(%)

模型	*Precision*	*Recall*	*F*1
RoGCN-ATT	**79.85**	**76.05**	**78.56**
-Stanza	64.60	66.71	65.27
-GloVe	73.36	76.05	74.22
-no-ATT	78.05	74.32	76.78

结束语 本文基于 RoBERTa 中文预训练模型，结合图卷积神经网络提出了一种用于中文地质领域的实体关系抽取模型 RoGCN-ATT，通过结合句子语义信息和结构信息，该模型能够更好地捕捉地质实体之间的复杂关系，提高了实体关系抽取的准确性和效率。利用 RoBERTa 中文预训练模型结合 BiLSTM 进行序列编码，使用注意力机制构建逻辑邻接矩阵，结合加权图卷积网络进行依赖信息编码，在构建的地质数据集上证明了本文模型的优越性。由于数据被处理为适用于图结构的形式，因此会存在不完善的信息，对实验结果产生影响。后续工作会在本文改进模型的基础上结合迁移学习等方法来提升模型的实体关系抽取性能，使其更加适用于地质领域，同时可以探索将该模型应用于其他领域的可能性。

参 考 文 献

[1] LI C, LIU D, ZHOU D, et al. Application and Prospect of Artificial Intelligence in the Field of Geology[J]. Bulletin of Mineralogy, Petrology and Geochemistry, 2022, 41(3): 668-677.

[2] MA R X. Research on Key Technologies of Knowledge Graph Construction in Chinese Medical Field[D]. Hangzhou: Zhejiang University, 2023.

[3] LI X, GAO R, QIN H, et al. EINE: Relation Classification by Enhancing the Impact of Non-Entity words[C]// Proceedings of the 2022 5th International Conference on Machine Learning and Natural Language Processing. 2022: 68-73.

[4] GUO Q, SUN Y, LIU G, et al. Constructing Chinese historical literature knowledge graph based on BERT[C]// Web Information Systems and Applications: 18th International Conference, WISA 2021, Kaifeng, China, September 24-26, 2021, Proceedings 18. Springer International Publishing, 2021: 323-334.

[5] HUANG S B, SUN X W, LI R S. Relation Classification Method Based on Cross-sentence Contextual Information for Neural Network[J]. Computer Science, 2022, 49(S1): 119-124.

[6] EBERTS M, ULGES A. An End-to-end Model for Entity-level Relation Extraction using Multi-instance Learning[C]// Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. 2021: 3650-3660.

[7] LI Z, FU L. A Relation-Aware Span-Level Transformer Network for Joint Entity and Relation Extraction[C]// 2022 International Joint Conference on Neural Networks (IJCNN). IEEE, 2022: 1-8.

[8] LI H, HOU S L, TONG Q, et al. Entity Relation Extraction Method in Weapon Field Based on DCNN and GLU[J]. Computer Science, 2023, 50(6A): 220200112-7.

[9] YU X S, LI L Y, ZHOU J L, et al. AM FRel: A method for joint extraction of entity relations in Chinese electronic medical records[J]. Journal of Chongqing University of Technology (Natural Science), 2024, 38(2): 189-197.

[10] ZHANG J L, ZHANG Y F, WANG M Q, et al. Joint extraction of Chinese entity relations based on graph convolutional neural network[J]. Computer Engineering, 2021, 47(12): 103-111.

[11] CUI Y, CHE W, LIU T, et al. Revisiting PreTrained Models for Chinese Natural Language Processing[C]// Findings of the Association for Computational Linguistics: EMNLP 2020. 2020: 657-668.

[12] ZHANG S, ZHENG D, HU X, et al. Bidirectional long short-term memory networks for relation classification[C]// Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation. 2015: 73-78.

[13] KUMAR S. A survey of deep learning methods for relation extraction[J]. arXiv: 1705.03645, 2017.

[14] ZENG D, LIU K, LAI S, et al. Relation classification via convolutional deep neural network[C]// Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. 2014: 2335-2344.

[15] TAKASE S, OKAZAKI N, INUI K. Modeling semantic compositionality of relational patterns[J]. Engineering Applications of Artificial Intelligence, 2016, 50: 256-264.

[16] NASAR Z, JAFFRY S W, MALIK M K. Named entit-y recognition and relation extraction: State-of-the-art[J]. ACM Compu-

ting Surveys(CSUR),2021,54(1):1-39.
[17] LEI X,SONG W,FAN R,et al. Semi-supervised geological disasters named entity recognition using few labeled data[J]. Geo-Informatica,2023,27:263-288.
[18] FAN R,WANG L,YAN J,et al. Deep learning-based named entity recognition and knowledge graph construction for geological hazards[J]. ISPRS International Journal of Geo-Information,2019,9(1):1-22.
[19] LUO X,ZHOU W,WANG W,et al. Attention-based relation extraction with bidirectional gated recurrent unit and highway network in the analysis of geological data[J]. IEEE Access,2017,6:5705-5715.
[20] HUANG X S,ZHU Y Q,FU L J,et al. Research on a geological entity relation extraction model for gold mine based on BERT [J]. Journal of Geomechanics,2021,27(3):391-399.
[21] CHEN Z L,YUAN F,LI X H,et al. Based on BERT-BiLSTM-CRF model the named entity and relation joint extration of Chinese lithological description corpus[J]. Geological Review,2022,68(2):742-750.
[22] WANG Z G,WEN H Y,LU Q,et al. Joint extraction of open entity relation in geological field[J]. Computer Engineering and Design,2021,42(4):996-1005.
[23] WU X Y,DUAN Y X,CHANG L J,et al. Research on entity and relation joint extraction for geological domain[J]. Computer Engineering,2023,49(3):121-127.
[24] BUNESCU R,MOONEY R. A shortest path dependency kernel for relation extraction[C]// Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing. 2005:724-731.
[25] CAI R,ZHANG X,WANG H. Bidirectional recurrent convolutional neural network for relation classification[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics(Volume 1:Long Papers). 2016:756-765.
[26] HENDRICKX I,KIM S N,KOZAREVA Z,et al. SemEval-2010 Task 8:Multi-Way Classification of Semantic Relations between Pairs of Nominals[C]// Proceedings of the 5th International Workshop on Semantic Evaluation. 2010:33-38.
[27] ZHANG Y,QI P,MANNING C D. Graph Convolution over Pruned Dependency Trees Improves Relation Extraction [C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018:2205-2215.
[28] YU B,MENGGE X,ZHANG Z,et al. Learning to prune dependency trees with rethinking for neural relation extraction [C]// Proceedings of the 28th International Conference on Computational Linguistics. 2020:3842-3852.
[29] HONG Y,LIU Y,YANG S,et al. Improving graph convolutional networks based on relation-aware attention for end-to-end relation extraction[J]. IEEE Access,2020,8:51315-51323.
[30] TIAN Y,CHEN G,SONG Y,et al. Dependency-driven relation extraction with attentive graph convolutional networks[C]// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing(Volume 1:Long Papers). 2021:4458-4471.
[31] ZHOU H,XU Y,YAO W,et al. Global context enhanced graph convolutional networks for document-level relation extraction [C]// Proceedings of the 28th International Conference on Computational Linguistics. 2020:5259-5270.
[32] DUAN J Y,YANG X,WANG H,et al. Document-level Relation Extraction of Graph Attention Convolutional Network Based on Inter-sentence Information [J]. Computer Science, 2023, 50(S1):220800189-6.
[33] ZHAO K,XU H,CHENG Y,et al. Representation iterative fusion based on heterogeneous graph neural network for joint entity and relation extraction[J]. Knowledge-Based Systems,2021, 219:106888.
[34] ZHOU L,WANG T,QU H,et al. A weighted GCN with logical adjacency matrix for relation extraction[M]// ECAI 2020. IOS Press,2020:2314-2321.
[35] PENNINGTON J,SOCHER R,MANNING C D. Glove:Global vectors for word representation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing(EMNLP). 2014:1532-1543.
[36] LI S,ZHAO Z,HU R,et al. Analogical Reasoning on Chinese Morphological and Semantic Relations[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics(Volume 2:Short Papers). 2018:138-143.
[37] HE H,CHOI J D. The Stem Cell Hypothesis:Dilemma behind Multi-Task Learning with Transformer Encoders[C]// Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021:5555-5577.
[38] LIN Y,SHEN S,LIU Z,et al. Neural relation extraction with selective attention over instances[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics(Volume 1:Long Papers). 2016:2124-2133.
[39] MANDYA A,BOLLEGALA D,COENEN F. Graph Convolution over Multiple Dependency Subgraphs for Relation Extraction[C]// COLING. International Committee on Computational Linguistics. 2020:6424-6435.
[40] QI P,ZHANG Y,ZHANG Y,et al. Stanza:A Python Natural Language Processing Toolkit for Many Human Languages [C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2020:101-108.

ZHANG Lu, born in 1999, postgraduate, is a member of CCF (No. I1760G). Her main research interests include knowledge graph, relation extraction, and so on.

DUAN Youxiang, born in 1964, Ph.D, professor, is a member of CCF (No. 05290S). His main research interests include network and service computing, the application of computer technology in oil and gas field, and so on.

（责任编辑：喻藜）